Meta Gene 4 (2015) 118–128

Contents lists available at ScienceDirect

Meta Gene

Initial description of primate-specific cystine-knot Prometheus genes and differential gene expansions of D-dopachrome tautomerase genes

Marko Premzl

Laboratory of Genomics, Centre of Animal Reproduction, 55 Heinzel St., Zagreb 10000, Croatia

ARTICLE INFO

Article history:
Received 3 December 2014
Revised 10 January 2015
Accepted 9 February 2015
Available online 25 April 2015

Keywords:
Comparative genomic analysis
Gene annotations
Molecular evolution
Phylogenetic analysis

ABSTRACT

Using eutherian comparative genomic analysis protocol and public genomic sequence data sets, the present work attempted to update and revise two gene data sets. The most comprehensive third party annotation gene data sets of eutherian adenohypophysis cystine-knot genes (128 complete coding sequences), and D-dopachrome tautomerases and macrophage migration inhibitory factor genes (30 complete coding sequences) were annotated. For example, the present study first described primate-specific cystine-knot Prometheus genes, as well as differential gene expansions of D-dopachrome tautomerase genes. Furthermore, new frameworks of future experiments of two eutherian gene data sets were proposed.

Introduction

The free availability of public eutherian genomic sequence data sets ushered in new era in field of eutherian comparative genomics (Flicek et al., 2014; Margulies et al., 2005; Murphy et al., 2001; O'Leary et al., 2013; Wilson and Reeder, 2005). Indeed, the public eutherian genomic sequences were suitable in computational analyses (Blakesley et al., 2004; Lindblad-Toh et al., 2011). For example, new human gene annotations were expected to revise gene data sets (Harrow et al., 2012; International Human Genome Sequencing Consortium, 2001). In biomedical research, such new human gene annotations were expected to uncover potential new drugs and drug targets, as well as to contribute to better understanding of both physiological and pathological processes. For example, the comprehensive eutherian adenohypophysis cystine-knot gene data sets included major protein hormone genes of both clinical and physiological importance, such as thyroid

E-mail address: Marko.Premzl@alumni.anu.edu.au.

http://dx.doi.org/10.1016/j.mgene.2015.02.005

stimulating hormone beta subunit genes, follicle stimulating hormone beta subunit genes, luteinizing hormone beta subunit genes and chorionic gonadotropin genes (Alvarez et al., 2009; Dos Santos et al., 2011; Jiang et al., 2014; Li and Ford, 1998; Roch and Sherwood, 2014). Likewise, the comprehensive eutherian D-dopachrome tautomerase and macrophage migration inhibitory factor gene data sets included key regulatory immune response genes (Esumi et al., 1998; Merk et al., 2011). Yet, because of the incompleteness of eutherian genomic sequence assemblies (Harrow et al., 2012) and potential sequence errors (International Human Genome Sequencing Consortium, 2004; Mouse Genome Sequencing Consortium, 2009) eutherian gene data sets were subject to future updates. For example, the eutherian comparative genomic analysis protocol was proposed as guidance in protection against sequence errors in public eutherian genomic sequence assemblies (Premzl, 2014a, 2014b, 2014c). The protocol included new test of reliability of public eutherian genomic sequences that used genomic sequence redundancies, as well as protein molecular evolution test that used relative synonymous codon usage statistics. Thus, using public eutherian genomic sequence data sets and new genomics and protein molecular evolution tests, the present work made attempts to update and revise gene data sets of eutherian adenohypophysis cystine-knot genes, and eutherian D-dopachrome tautomerase and macrophage migration inhibitory factor genes respectively.

Materials and methods

Gene annotations

The gene annotations included identification of genes in eutherian genomic sequence assemblies, analysis of gene features, tests of reliability of eutherian public genomic sequences and alignments of genomic sequences. The protocol made use of free available genomic sequence data sets in public databases and software. The BioEdit 7.0.5.3 program was used in nucleotide and protein sequence analyses (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The Ensembl genome browser, and its BLAST or BLAT tools were used in identification of genes in genomic sequence assemblies (http://www.ensembl.org/index.html) (Flicek et al., 2014). The analysis of gene features used direct evidence of eutherian gene annotations in NCBI's nr, est_human, est_mouse and est_others databases (http://www.ncbi.nlm.nih.gov). The protocol first annotated potential coding sequences that were tested using tests of reliability of eutherian public genomic sequences. The tests made use of genomic sequence redundancies and primary experimental sequence data in NCBI's Trace Archive database (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). The first test step included analysis of nucleotide sequence coverage of each potential coding sequence using primary experimental sequence data and NCBI's program Netblast (ftp://ftp.ncbi.nlm.nih.gov/blast/documents/netblast.html). The second test step included classification of potential coding sequences. The potential coding sequences were designated as complete coding sequences if consensus trace sequence coverage was available for every nucleotide. Alternatively, they were described as putative coding sequences. The complete coding sequences were used in phylogenetic and protein molecular evolution analyses. The guidelines of human and mouse gene nomenclature were used in gene descriptions (http://www.genenames.org/guidelines.html and http://www.informatics.jax.org/mgihome/nomen/gene.shtml). The complete coding sequence data sets were reviewed by EBI as third party annotation gene data sets (http://www.ebi.ac.uk/embl/Documentation/third_party_annotation_dataset.html). The alignments of genomic sequences first included identification and masking of transposable elements in genomic sequences. The RepeatMasker program version open-3.3.0 was used, using default settings except simple repeats and low complexity elements were not masked (sensitive mode, cross_match version 1.080812, RepBase Update 20110920, RM database version 20110920 (http://www.repeatmasker.org/). However, the *PREA1-3* and *PREB* genomic sequences were not masked in alignments. Then the mVISTA web tool was used in alignments of genomic sequences (http://genome.lbl.gov/vista/index.shtml). The default settings and AVID algorithm were used in pairwise alignments. The cut-offs of detection of common genomic sequence regions in each pairwise genomic sequence alignment were determined empirically (Supplementary data files 3 and 9). The potential regulatory genomic sequence regions were aligned using ClustalW implemented in BioEdit 7.0.5.3, and nucleotide sequence alignments were corrected manually. Using BioEdit 7.0.5.3, the pairwise nucleotide sequence identities of common predicted promoter genomic sequence regions were calculated and used in statistical analysis (Microsoft Office Excel).

Phylogenetic analysis

The phylogenetic analysis included alignments of protein sequences, alignments of nucleotide sequences, calculations of phylogenetic trees and calculations of nucleotide sequence identities. The complete coding sequences were first aligned at amino acid level using ClustalW implemented in BioEdit 7.0.5.3. Then the protein sequence alignments and nucleotide sequence alignments were corrected manually. The MEGA5 program was used in calculations of phylogenetic trees (http://www.megasoftware.net). The phylogenetic trees were calculated using neighbour-joining method (default settings, except gaps/missing data = pairwise deletion) (not shown), minimum evolution method (default settings, except gaps/missing data = pairwise deletion) and maximum parsimony method (default settings, except gaps/missing data = use all sites) (not shown). However, because their homogeneity and stationarity assumptions were not satisfied, the maximum likelihood methods were not used in phylogenetic analysis (data not shown). The pairwise nucleotide sequence identities of complete coding sequences were calculated using BioEdit 7.0.5.3. The calculations were used in statistical analysis (Microsoft Office Excel).

Protein molecular evolution analysis

The protein molecular evolution analysis included new tests of protein molecular evolution. The tests integrated patterns of nucleotide sequence similarities of aligned complete coding sequences with protein tertiary structures. The relative synonymous codon usage statistic *R* was calculated using MEGA5 as ratio between observed and expected amino acid codon counts. The amino acid codons with $R \leq 0.7$ were designated as not preferable amino acid codons. The not preferable amino acid codons in analysis of eutherian *GPB5*, *TSHB*, *FSHB* and *LHB-CGB* close gene homologues were: TTT (0.69), TTA (0.03), TTG (0.38), CTA (0.17), ATT (0.46), ATA (0.52), GTT (0.37), GTA (0.38), TCA (0.27), TCG (0.12), CCG (0.44), ACA (0.59), ACG (0.33), GCG (0.18), CAA (0.46), AAT (0.63), AAA (0.66), GAT (0.59), GAA (0.62), CGT (0.26), CGA (0.57), AGT (0.61), GGT (0.36) and GGA (0.64). In analysis of *GPA1* and *GPA2* close gene homologues, they were: TTA (0.06), TTG (0.52), CTA (0.23), ATT (0.53), ATA (0.52), GTT (0.37), GTA (0.39), TCA (0.34), TCG (0.16), CCG (0.38), ACG (0.35), GCG (0.25), CAA (0.41), AAT (0.63), GAT (0.7), GAA (0.61), CGT (0.28), CGA (0.53), GGT (0.51) and GGA (0.68). In analysis of *DDT* and *MIF* close gene homologues, the not preferable amino acid codons were: TTT (0.62), TTA (0.2), TTG (0.61), CTT (0.17), CTA (0.37), ATT (0.5), ATA (0.46), GTT (0.28), GTA (0.44), TCT (0.22), TCA (0.13), TCG (0.2), CCT (0.53), CCA (0.26), ACT (0.68), ACA (0.44), GCT = (0.4), GCA (0.29), TAT (0.42), CAT (0.18), CAA (0.09), AAT (0.23), AAA (0.64), GAT (0.19), GAA (0.1), TGT (0.35), CGT (0.14), (AGT (0.29), AGA (0.14), AGG (0.55), GGT (0.37) and GGA (0.15). In protein molecular evolution analyses, the reference protein sequence residues were designated as invariant amino acid sites (invariant alignment positions), forward amino acid sites (variant alignment positions that did not include amino acid codons with $R \leq 0.7$) or compensatory amino acid sites (variant alignment positions including amino acid codons with $R \leq 0.7$). Thus, the presence of preferable amino acid codons and absence of not preferable amino acid codons indicated that forward amino acid sites could have major influence on protein function. Conversely, the presence of not preferable amino acid codons indicated that compensatory amino acid sites could have minor influence on protein function. The DeepView/Swiss-PdbViever 4.0.1 program was used for analyses of protein tertiary structures (http://spdbv.vital-it.ch/). The prediction of N-terminal signal peptide presence was undertaken using SignalP-4.0 (http://www.cbs.dtu.dk/services/SignalP/).

Results and discussion

Initial description of primate-specific cystine-knot Prometheus genes

Gene annotations

The present analysis annotated most comprehensive data set of eutherian adenohypophysis cystine-knot genes. Among 183 potential coding sequences, the comparative genomic analysis protocol annotated 128 complete coding sequences encoding 11 Prometheus proteins (PREA1-3 and PREB), 23 glycoprotein-B5 proteins (GPB5), 19 thyroid stimulating hormone beta subunits (TSHB), 21 follicle stimulating hormone beta subunits (FSHB), 19 luteinizing hormone beta subunits and chorionic gonadotropins (LHB and CGB), 18 glycoprotein hormone alpha subunits (GPA1) and 17 glycoprotein-A2 proteins (GPA2)

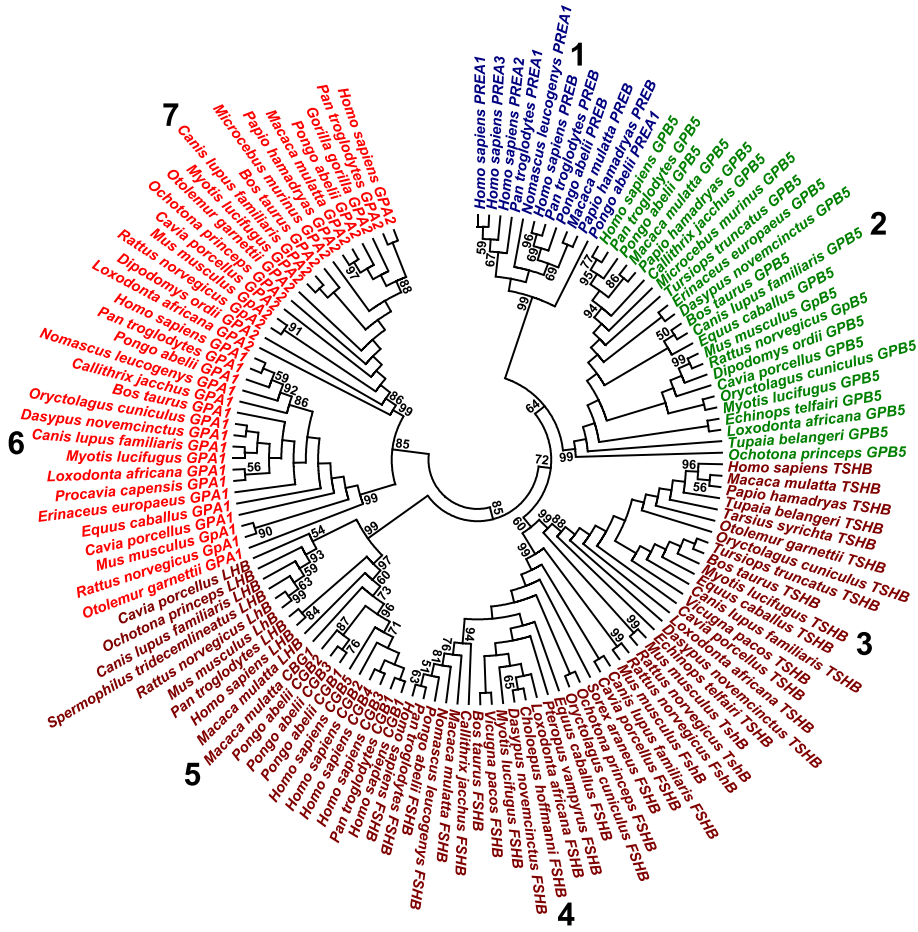


Fig. 1. Phylogenetic analysis of eutherian adenohypophysis cystine-knot genes. The minimum evolution tree of eutherian adenohypophysis cystine-knot genes was calculated using maximum composite likelihood method. The estimates >50% were shown, after 1000 bootstrap replicates. The major gene clusters were indicated using numbers (1–7).

(Fig. 1). The gene data set was made available in public databases as one third party annotation gene data set (http://www.ebi.ac.uk/ena/data/view/HF564658-HF564785) (Supplementary data file 1). The present work integrated gene annotations, phylogenetic analysis, and protein molecular evolution analysis and first described primate-specific cystine-knot genes that were named Prometheus genes (Figs. 1 and 2). Whereas the *PREA1-3* genes were annotated in Hominidae genomic sequence assemblies, *PREB* gene was annotated in Hominidae and Cercopithecidae genomic sequence assemblies. There were both direct and indirect evidence of *PREA1-3* and *PREB* gene annotations (Clamp et al., 2007). The direct evidence included gene transcripts (Supplementary data file 2). For example, the present *PREA1-3* and *PREB* transcript data set annotated seven human *PREA2* gene exons, eight human *PREA3* gene exons and nine human *PREB* gene exons. However, the annotated primate-specific *PREA1-3* and *PREB* ORFs were encoded by single translated exons (Supplementary data file 3A). Next, in present primate genomic sequence assemblies, the *PREA1-3* and *PREB* genes were positioned within segmental duplications on chromosome 17 along minimally ~31–35 kb that showed >90% nucleotide sequence identities (Supplementary data file 3A). For example, the human *PREA2* and *PREA3* genes were positioned within segmental

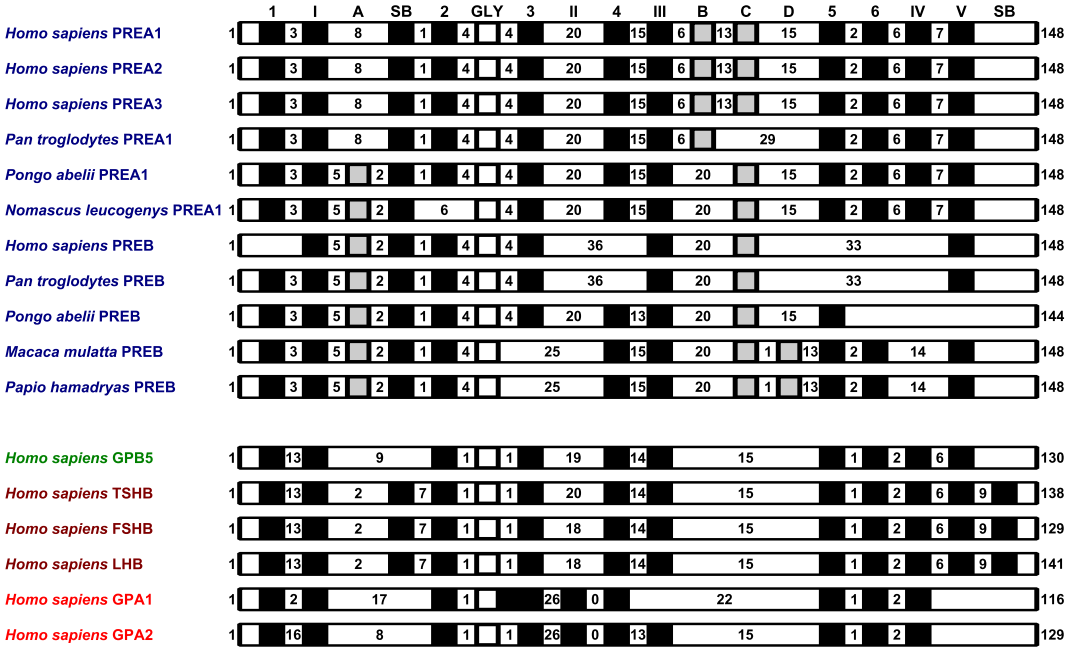


Fig. 2. Distribution of cysteines in cystine-knot domains of human adenohypophysis cystine-knot proteins. The black rectangles indicated common cysteine residues. The grey rectangles indicated cysteines in primate-specific PREA1-3 and PREB proteins. The white rectangles indicated glycine residues in amino acid motifs Cys-4x-Gly-4x-Cys or Cys-1x-Gly-1x-Cys. The numbers between rectangles indicated numbers of amino acids. Whereas the common cysteine residues were labelled according to Alvarez et al. (2009) and present analysis, PREA1-3- and PREB-specific cysteine residues were labelled A–D.

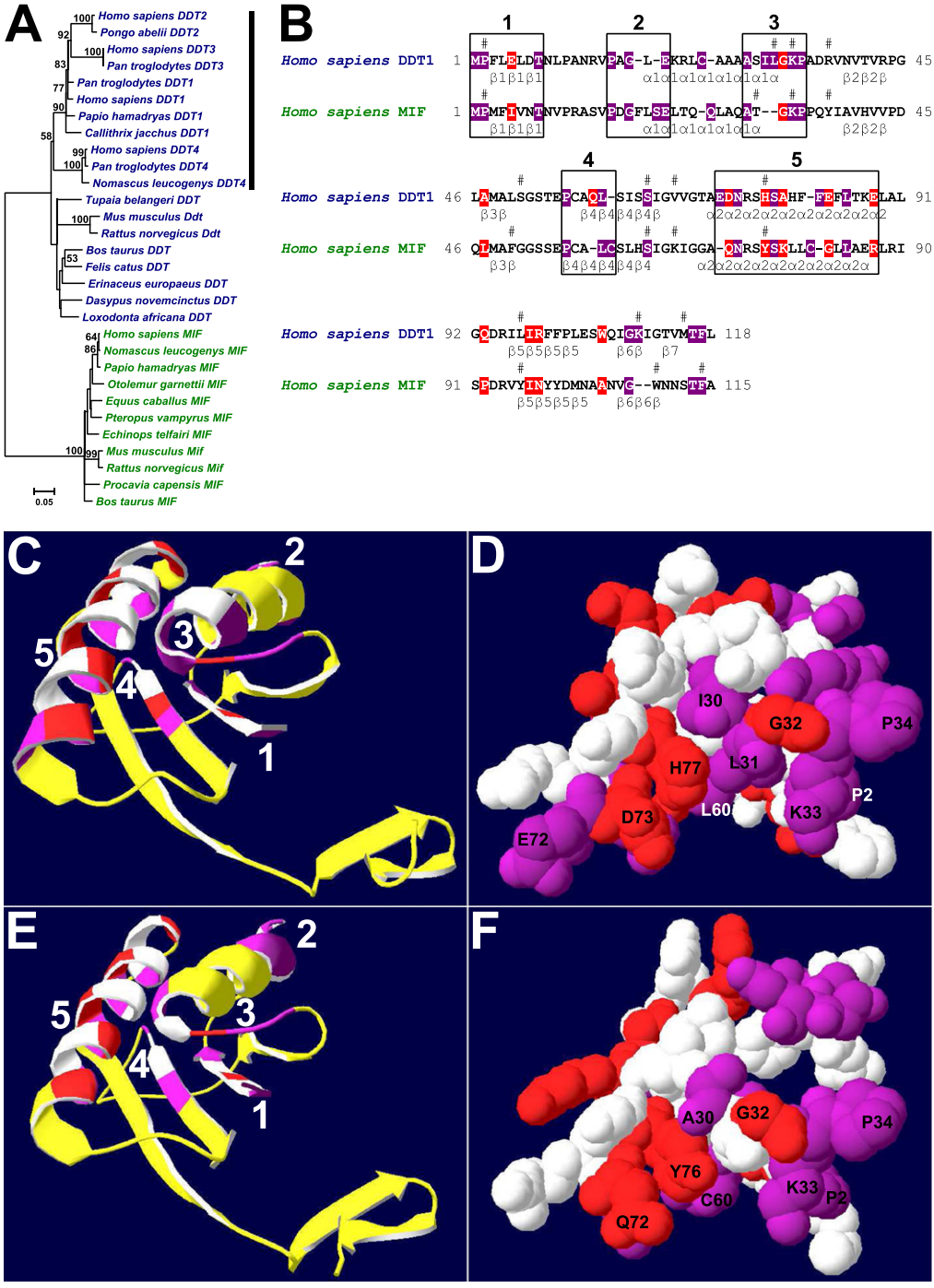
duplications along >243 kb. Likewise, the primate-specific *PREA1-3* and *PREB* ORFs showed nucleotide sequence similarities to transposable elements. For example, within human *PREA1* ORF there were nucleotide sequence similarities to transposable elements MIR3 (151–294 bp) and L2c (359–447 bp). Finally, the computational gene annotations of new adenohypophysis cystine-knot genes were known in human *GPB5* and *GPA2* genes (Hsu et al., 2002; Macdonald et al., 2005).

The present analysis annotated both new and known potential regulatory genomic sequence regions of eutherian adenohypophysis cystine-knot genes, as guidelines of future experiments. In eutherian *GPB5* promoters, there were two common genomic sequence regions that showed nucleotide sequence identity patterns that exceeded criteria of detection of potential regulatory genomic sequence regions (Supplementary data file 3B). In eutherian *TSHB* promoters, there was one common potential regulatory genomic sequence region (Supplementary data file 3C, Supplementary data file 4A). For example, the common potential regulatory genomic sequence region included functional PIT1 and GATA2 *cis*-elements, suppressor region and TATA *cis*-element (Kashiwabara et al., 2009). The average pairwise nucleotide sequence identity of common potential regulatory genomic sequence region was $\bar{a} = 0.847$ ($a_{max} = 0.996$, $a_{min} = 0.745$, $\bar{a}_{ad} = 0.034$). There was one common potential regulatory genomic sequence region in eutherian *FSHB* promoters (Supplementary data file 3D, Supplementary data file 4B). For example, the common potential regulatory genomic sequence region included functional LHX3 *cis*-elements A-C and TATA *cis*-element (West et al., 2004). The average pairwise nucleotide sequence identity of common potential regulatory genomic sequence region was $\bar{a} = 0.787$ ($a_{max} = 0.992$, $a_{min} = 0.622$, $\bar{a}_{ad} = 0.073$). In eutherian *LHB-CGB* promoters, there were four common potential regulatory genomic sequence regions (Supplementary data file 3E, Supplementary data file 4C). For example, the common potential regulatory genomic sequence region 4 included functional SF-1 and Egr-1 *cis*-elements and TATA *cis*-element (Horton and Halvorson, 2004). The average pairwise nucleotide sequence identity of common potential regulatory genomic sequence region 4 was $\bar{a} = 0.77$ ($a_{max} = 0.986$, $a_{min} = 0.461$, $\bar{a}_{ad} = 0.095$). There were four common potential regulatory genomic sequence regions in eutherian *GPA1* promoters (Supplementary data file 3F, Supplementary data file 4D). For example, the common potential regulatory genomic sequence region 4 included functional GSE *cis*-element, CRE *cis*-element, Hominidae-specific CRE *cis*-element and TATA *cis*-element (Fowkes et al., 2003). The average pairwise nucleotide sequence identity of common potential regulatory genomic sequence region 4 was $\bar{a} = 0.766$ ($a_{max} = 0.983$, $a_{min} = 0.602$, $\bar{a}_{ad} = 0.06$). Finally, in eutherian *GPA2* promoters, there were two common potential regulatory genomic sequence regions (Supplementary data file 3G).

Phylogenetic analysis

The present analysis described seven major gene clusters of eutherian adenohypophysis cystine-knot genes (Fig. 1). The major gene cluster 1 included primate-specific *PREA1-3* and *PREB* genes. The eutherian *GPB5* genes comprised major gene cluster 2. The eutherian *TSHB* genes, *FSHB* genes and *LHB-CGB* genes were grouped in major gene clusters 3, 4 and 5 respectively (Li and Ford, 1998). Whereas the eutherian *GPA1* genes were included in major gene cluster 6, major cluster 7 was comprised of eutherian *GPA2* genes. The identical major tree branching patterns were calculated using minimum evolution (Fig. 1), neighbour-joining and maximum parsimony methods. The present eutherian adenohypophysis cystine-knot gene classification was confirmed by calculations of nucleotide sequence identity patterns (Supplementary data file 5). The major gene cluster 1 paralogues showed high nucleotide sequence identities that were typical in primate-specific gene expansions. In comparisons with major gene cluster 2 genes, the major gene cluster 1 genes showed nucleotide sequence identity patterns of typical homologues, but in comparisons with other major gene clusters, major gene cluster 1 genes showed nucleotide sequence identity patterns of distant homologues. The eutherian major gene cluster 2-7 genes respectively showed nucleotide sequence identities that were typical in comparisons between eutherian orthologues. In comparisons between major gene cluster 2-5 genes, there were nucleotide sequence identity patterns of close homologues, as well as in comparisons between major gene cluster 6 and 7 genes. Finally, in comparisons of major gene cluster 2-5 genes with major gene cluster 6 and 7 genes, there were nucleotide sequence identity patterns of typical homologues. The exceptions were nucleotide sequence identity patterns of distant homologues between major gene cluster 3 and 6 genes. The present grouping of eutherian *LHB-CGB* genes into one major gene cluster (Lapthorn et al., 1994; Li and Ford, 1998) was different to analysis of Hsu et al. (Hsu et al., 2002) that grouped *LHB* and *CGB* genes in two groups. Indeed, the eutherian *LHB* genes and primate-specific *CGB* genes showed nucleotide sequence

identity patterns that were typical in comparisons of eutherian orthologues and paralogues. Finally, the eutherian *LHB* genes and primate-specific *CGB* genes included common potential regulatory genomic sequence regions (Supplementary data file 3E, Supplementary data file 4C).



Protein molecular evolution analysis

In cystine-knot domains, the PREA1-3 and PREB proteins included 8–13 common cysteines and PREA1-3- and PREB-specific cysteines, GPB5 proteins included 10 common cysteines, TSHB, FSHB and LHB and CGB proteins included 11 common cysteines and GPA1 and GPA2 proteins included 10 common cysteines (Fig. 2). The primate-specific PREA1-3 and PREB proteins included new cystine-knot domain cysteine patterns. For example, the common Cys(2)-1x-Gly-1x-Cys(3) amino acid sequence motif (Alvarez et al., 2009; Lapthorn et al., 1994) was replaced by Cys(2)-4x-Gly-4x-Cys(3) amino acid sequence motif. Likewise, the common Cys(5)-1x-Cys(6) amino acid sequence motif (Alvarez et al., 2009; Lapthorn et al., 1994) was replaced by Cys(5)-2x-Cys(6) amino acid sequence motif. However, one major difference between PREA1-3 and PREB proteins and adenohypophysis cystine-knot homologues was no prediction of N-terminal signal peptides in primate-specific PREA1-3 and PREB proteins by SignalP analysis. In addition, they did not include potential N-glycosylation signal sites.

The new tests of protein molecular evolution were used in molecular evolution analysis of eutherian GPB5, TSHB, FSHB and LHB-CGB close protein homologues, including 82 complete coding sequences (Supplementary data file 5, Supplementary data file S6). The human FSHB was used as reference protein amino acid sequence in analysis of FSHB crystal structure 1XWD (Fan and Hendrickson, 2005; Lapthorn et al., 1994). In human FSHB protein amino acid sequence, there were 16 invariant amino acid sites and 7 forward amino acid sites (Supplementary data file 7A, Supplementary data file 7C–D). The present analysis described two amino acid clusters with overrepresented invariant and/or forward amino acid sites that were positioned between amino acid positions W45-R53 and A97-C105. The amino acid clusters included common amino acid sequence motifs Cys(2)-1x-Gly-1x-Cys(3) (Cluster 1) and Cys(5)-1x-Cys(6) (Cluster 2) (Fan and Hendrickson, 2005; Lapthorn et al., 1994). The new tests of protein molecular evolution were used in molecular evolution analysis of eutherian GPA1 and GPA2 close protein homologues including 35 complete coding sequences (Supplementary data file 5, Supplementary data file S6). Using human GPA1 as reference protein amino acid sequence and crystal structure 1XWD (Fan and Hendrickson, 2005; Lapthorn et al., 1994), the present analysis described 27 invariant amino acid sites and 26 forward amino acid sites (Supplementary data file 7B, Supplementary data file 7E–F). There were five amino acid clusters with overrepresented invariant and/or forward amino acid sites that were positioned between amino acid positions: C31-Q37, F42-T63, K75-S79, C83-G96 and E101-K115. The amino acid cluster 1 included amino acid site Cys(I), amino acid cluster 2 included common amino acid sequence motif Cys(2)-1x-Gly-1x-Cys(3), amino acid cluster 4 included amino acid site Cys(4) and amino acid cluster 5 included common amino acid sequence motif Cys(5)-1x-Cys(6) (Fan and Hendrickson, 2005; Lapthorn et al., 1994).

Initial description of primate-specific differential gene expansions of D-dopachrome tautomerase genes

Gene annotations

Among 49 potential coding sequences, the eutherian comparative genomic analysis protocol annotated 30 complete coding sequences of 19 eutherian D-dopachrome tautomerases (DDT) and 11 eutherian migration inhibitory factors (MIF) (Fig. 3A). The present most comprehensive eutherian *DDT* and *MIF* gene data set was made available in public databases as one third party annotation gene data set (http://www.ebi.ac.uk/ena/

Fig. 3. Analysis of eutherian D-dopachrome tautomerase and macrophage migration inhibitory factor genes. (A) Minimum evolution tree of eutherian D-dopachrome tautomerase and macrophage migration inhibitory factor genes. The tree was calculated using maximum composite likelihood method. The estimates >50% were shown, after 1000 bootstrap replicates. The vertical bar labelled primate-specific differential gene expansions *DDT1-4*. (B–F) Protein molecular evolution analysis. (B) Reference human DDT1 and MIF protein amino acid sequences. The invariant amino acid sites were shown using white letters on violet backgrounds. The forward amino acid sites were shown using white letters on red backgrounds. The amino acid clusters 1–5 with overrepresented invariant or/and forward amino acid sites were labelled using rectangles. The secondary structure elements were designated according to Sugimoto et al. (1999) for DDT1 crystal structure 1DPT, and according to Sun et al. (1996) for MIF crystal structure 1MIF. The amino acid residues implicated in putative DDT1 and MIF active sites (Sugimoto et al., 1999) were labelled by #s. (C–D) Analysis of human DDT1 crystal structure 1DPT. (C) Ribbon representation of human DDT1 crystal structure 1DPT. (D) van-der-Waals representations of amino acid cluster 1-5 amino acids in view identical to that in C. (E–F) Analysis of human MIF crystal structure 1MIF. (E) Ribbon representation of human MIF crystal structure 1MIF. (F) van-der-Waals representations of amino acid cluster 1-5 amino acids in view identical to that in E. (C–F) The amino acid cluster 1-5 invariant amino acid sites were labelled violet, forward amino acid sites were labelled red and compensatory amino acid sites were labelled white.

data/view/HF564786-HF564815) (Supplementary data file 8). The present work integrated gene annotations, phylogenetic analysis and protein molecular evolution analysis and first described primate-specific differential gene expansions *DDT1-4* (Fig. 3A). For example, the human *DDT1-4* genes were present in Hominidae genomic sequence assemblies. There were direct and indirect evidence of gene annotations of primate-specific differential gene expansions *DDT1-4* (Clamp et al., 2007). The direct evidence included gene transcripts (Supplementary data file 2). The indirect evidence included nucleotide sequence identity patterns of complete coding sequences, untranslated genomic sequence regions, promoters and introns (Supplementary data file 5, Supplementary data file 9). For example, the human *DDT1* exon 1 translated genomic sequence and exon 2 each showed 100% nucleotide sequence identities in comparisons with human *DDT2* exon 1 translated genomic sequence and exon 2. In addition, the primate-specific *DDT2* and *DDT3* genes showed nucleotide sequence similarities along human *DDT1* promoter and introns. The primate-specific *DDT4* genes showed nucleotide sequence similarities along human *DDT1* exon 3. The alignments of *DDT* genomic sequences indicated that there was one common potential regulatory genomic sequence region in *DDT* and *MIF* promoters, respectively (Supplementary data file 9).

Phylogenetic analysis

The minimum evolution (Fig. 3A), neighbour-joining and maximum parsimony phylogenetic trees of eutherian *DDT* and *MIF* genes showed similar topologies with identical major branching patterns that clustered *DDT* and *MIF* genes in two major gene clusters. Indeed, the primate-specific *DDT1-4* genes were grouped separately. The present calculations of nucleotide sequence identity patterns between major gene clusters described typical eutherian *DDT* and *MIF* genes as close homologues (Supplementary data file 5).

Protein molecular evolution analysis

The new tests of protein molecular evolution were used in molecular evolution analysis of eutherian DDT and MIF close protein homologues including 30 complete coding sequences (Fig. 3B-F, Supplementary data file 10). The present analysis determined 24 invariant amino acid sites and 13 forward amino acid sites in reference human DDT1 protein amino acid sequence, and 22 invariant amino acid sites and 12 forward amino acid sites in reference human MIF protein amino acid sequence. Whereas the amino acids implicated in putative DDT1 active site (Sugimoto et al., 1999) included five invariant amino acid sites and one forward amino acid site, amino acids implicated in putative MIF active site (Sugimoto et al., 1999) included four invariant amino acid sites and one forward amino acid site. There were five amino acid clusters with overrepresented invariant or/and forward amino acid sites determined in reference protein amino acid sequences: cluster 1 (M1-T8 in both DDT1 and MIF), cluster 2 (P16-E20 in DDT1 and P16-E22 in MIF), cluster 3 (A28-P34 in DDT1 and A30-P34 in MIF), cluster 4 (P56-L60 in DDT1 and P56-C60 in MIF) and cluster 5 (E72-E88 in DDT1 and Q72-R87 in MIF). The present analysis of human DDT1 and MIF crystal structures 1DPT and 1MIF indicated that amino acids in amino acid clusters 1 and 3, C-terminal part of amino acid cluster 4 and N-terminal part of amino acid cluster 5 delineated potential human DDT1 and MIF active sites (Sugimoto et al., 1999).

Conclusions

The present analysis updated and revised gene data sets of eutherian adenohypophysis cystine-knot genes, and *DDT* and *MIF* genes respectively. The most comprehensive third party annotation gene data sets were annotated. Indeed, the present study first described primate-specific cystine-knot *PREA1-3* and *PREB* genes, as well as differential gene expansions *DDT1-4*. The eutherian comparative genomic analysis protocol proposed new frameworks of future experiments of two eutherian gene data sets.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mgene.2015.02.005.

References

Alvarez, E., Cahoreau, C., Combarnous, Y., 2009. Comparative structure analyses of cystine knot-containing molecules with eight aminoacyl ring including glycoprotein hormones (GPH) alpha and beta subunits and GPH-related A2 (GPA2) and B5 (GPB5) molecules. Reprod. Biol. Endocrinol. 7, 90.

Blakesley, R.W., Hansen, N.F., Mullikin, J.C., Thomas, P.J., McDowell, J.C., Maskeri, B., Young, A.C., Benjamin, B., Brooks, S.Y., Coleman, B.I., Gupta, J., Ho, S.L., Karlins, E.M., Maduro, Q.L., Stantripop, S., Tsurgeon, C., Vogt, J.L., Walker, M.A., Masiello, C.A., Guan, X., Comparative Sequencing Program, N.I.S.C., Bouffard, G.G., Green, E.D., 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 14, 2235–2244.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., Lander, E.S., 2007. Distinguishing protein-coding and noncoding genes in the human genome. Proc. Natl. Acad. Sci. U. S. A. 104, 19428–19433.

Dos Santos, S., Mazan, S., Venkatesh, B., Cohen-Tannoudji, J., Quérat, B., 2011. Emergence and evolution of the glycoprotein hormone and neurotrophin gene families in vertebrates. BMC Evol. Biol. 11, 332.

Esumi, N., Budarf, M., Ciccarelli, L., Sellinger, B., Kozak, C.A., Wistow, G., 1998. Conserved gene structure and genomic linkage for D-dopachrome tautomerase (DDT) and MIF. Mamm. Genome 9, 753–757.

Fan, Q.R., Hendrickson, W.A., 2005. Structure of human follicle-stimulating hormone in complex with its receptor. Nature 433, 269–277.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., Searle, S.M., 2014. Ensembl 2014. Nucleic Acids Res. 42, D749–D755.

Fowkes, R.C., Desclozeaux, M., Patel, M.V., Aylwin, S.J., King, P., Ingraham, H.A., Burrin, J.M., 2003. Steroidogenic factor-1 and the gonadotrope-specific element enhance basal and pituitary adenylate cyclase-activating polypeptide-stimulated transcription of the human glycoprotein hormone alpha-subunit gene in gonadotropes. Mol. Endocrinol. 17, 2177–2188.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J., 2012. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774.

Horton, C.D., Halvorson, L.M., 2004. The cAMP signaling system regulates LHbeta gene expression: roles of early growth response protein-1, SP1 and steroidogenic factor-1. J. Mol. Endocrinol. 32, 291–306.

Hsu, S.Y., Nakabayashi, K., Bhalla, A., 2002. Evolution of glycoprotein hormone subunit genes in bilateral metazoa: identification of two novel human glycoprotein hormone subunit family genes, GPA2 and GPB5. Mol. Endocrinol. 16, 1538–1551.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

Jiang, Xuliang, Dias, James A., He, Xiaolin, 2014. Structural biology of glycoprotein hormones and their receptors: Insights to signaling. Mol. Cell. Endocrinol. 382, 424–451.

Kashiwabara, Y., Sasaki, S., Matsushita, A., Nagayama, K., Ohba, K., Iwaki, H., Matsunaga, H., Suzuki, S., Misawa, H., Ishizuka, K., Oki, Y., Nakamura, H., 2009. Functions of PIT1 in GATA2-dependent transactivation of the thyrotropin beta promoter. J. Mol. Endocrinol. 42, 225–237.

Lapthorn, A.J., Harris, D.C., Littlejohn, A., Lustbader, J.W., Canfield, R.E., Machin, K.J., Morgan, F.J., Isaacs, N.W., 1994. Crystal structure of human chorionic gonadotropin. Nature 369, 455–461.

Li, M.D., Ford, J.J., 1998. A comprehensive evolutionary analysis based on nucleotide and amino acid sequences of the alpha- and beta-subunits of glycoprotein hormone gene family. J. Endocrinol. 156, 529–542.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L.D., Lowe, C.B., Holloway, A.K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M.J., Jaffe, D.B., Jungreis, I., Kent, W.J., Kostka, D., Lara, M., Martins, A.L., Massingham, T., Moltke, I., Raney, B.J., Rasmussen, M.D., Robinson, J., Stark, A., Vilella, A.J., Wen, J., Xie, X., Zody, M.C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin, J., Bloom, T., Chin, C.W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K.C., Kovar, C.L., Muzny, D.M., Gibbs, R.A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree, A., Dihn, H.H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L.R., Nazareth, L.V., Okwuonu, G., Santibanez, J., Warren, W.C., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Genome Institute at Washington University, Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E.H., Herrero, J., Green, E.D., Haussler, D., Siepel, A., Goldman, N., Pollard, K.S., Pedersen, J.S., Lander, E.S., Kellis, M., 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482.

Macdonald, L.E., Wortley, K.E., Gowen, L.C., Anderson, K.D., Murray, J.D., Poueymirou, W.T., Simmons, M.V., Barber, D., Valenzuela, D.M., Economides, A.N., Wiegand, S.J., Yancopoulos, G.D., Sleeman, M.W., Murphy, A.J., 2005. Resistance to diet-induced obesity in mice globally overexpressing OGH/GPB5. Proc. Natl. Acad. Sci. U. S. A. 102, 2496–2501.

Margulies, E.H., Vinson, J.P., NISC Comparative Sequencing Program, Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., Clamp, M., 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc. Natl. Acad. Sci. U. S. A. 102, 4795–4800.

Merk, M., Zierow, S., Leng, L., Das, R., Du, X., Schulte, W., Fan, J., Lue, H., Chen, Y., Xiong, H., Chagnon, F., Bernhagen, J., Lolis, E., Mor, G., Lesur, O., Bucala, R., 2011. The D-dopachrome tautomerase (DDT) gene product is a cytokine and functional homolog of macrophage migration inhibitory factor (MIF). Proc. Natl. Acad. Sci. U. S. A. 108, E577–E585.

Mouse Genome Sequencing Consortium, 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 7 e1000112.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614–618.

O'Leary, M.A., Bloch, J.I., Flynn, J.J., Gaudin, T.J., Giallombardo, A., Giannini, N.P., Goldberg, S.L., Kraatz, B.P., Luo, Z.X., Meng, J., Ni, X., Novacek, M.J., Perini, F.A., Randall, Z.S., Rougier, G.W., Sargis, E.J., Silcox, M.T., Simmons, N.B., Spaulding, M., Velazco, P.M., Weksler, M., Wible, J.R., Cirranello, A.L., 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339, 662–667.

Premzl, M., 2014a. Comparative genomic analysis of eutherian Mas-related G protein-coupled receptor genes. Gene 540, 16–19.

Premzl, M., 2014b. Comparative genomic analysis of eutherian ribonuclease A genes. Mol. Genet. Genomics 289, 161–167.

Premzl, M., 2014c. Third party annotation gene data set of eutherian lysozyme genes. Genomics Data 2, 258–260.
Roch, G.J., Sherwood, N.M., 2014. Glycoprotein hormones and their receptors emerged at the origin of metazoans. Genome Biol. Evol. 6, 1466–1479.
Sugimoto, H., Taniguchi, M., Nakagawa, A., Tanaka, I., Suzuki, M., Nishihira, J., 1999. Crystal structure of human D-dopachrome tautomerase, a homologue of macrophage migration inhibitory factor, at 1.54 A resolution. Biochemistry 38, 3268–3279.
Sun, H.W., Bernhagen, J., Bucala, R., Lolis, E., 1996. Crystal structure at 2.6-A resolution of human macrophage migration inhibitory factor. Proc. Natl. Acad. Sci. U. S. A. 93, 5191–5196.
West, B.E., Parker, G.E., Savage, J.J., Kiratipranon, P., Toomey, K.S., Beach, L.R., Colvin, S.C., Sloop, K.W., Rhodes, S.J., 2004. Regulation of the follicle-stimulating hormone beta gene by the LHX3 LIM-homeodomain transcription factor. Endocrinology 145, 4866–4879.
Wilson, D.E., Reeder, D.M., 2005. Mammal Species of the World: A Taxonomic and Geographic Reference. 3rd edn. The Johns Hopkins University Press, Baltimore.